JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 24, 440–445 (1968)

On the Set of Optimal Policies in Discrete Dynamic Programming

STEVEN A. LIPPMAN

University of California, Los Angeles, California

Submitted by Richard Bellman

1. INTRODUCTION

We consider a finite state, finite action space Markovian decision process. One problem is to choose a policy, termed β-optimal, that maximizes the total discounted expected income for an infinite number of time periods. It is well known [1] that the set of β-optimal policies contains a stationary policy. We show that the set of β-optimal policies consists of all sequences of one period rules (mappings from the state space to the action space) such that if any rule in the sequence were to be used repeatedly as a stationary policy, it would be β-optimal. Apparently this result is intuitively obvious. However, when we seek a policy, termed optimal, to maximize the expected return per unit time, the analogous result is not true (see example 1). In this case we give a partial characterization of the set of optimal policies.

2. NOTATION

Following the notation of Blackwell [1], we make the following definitions: $\{1, 2,..., S\}$ is the set of states, A_s is the set of actions available in state s $(1 \leqslant s \leqslant S)$, $i(s, a)$ is the immediate income received from taking action a while in state s, and $q(s' \mid s, a)$ is the conditional probability that the system is in state s' at time $n + 1$, given that the system was in state s and that action a was taken at time n $(n = 1, 2,...)$.

Letting $F = \mathsf{X}_{s=1}^{S} A_s$, a *policy* π is a sequence $\langle f_i \rangle_{i=1}^{\infty}$ from F. Using the policy π means that action $f_n(s)$ is chosen if the system is in state s at time n. For $f \in F$, let $f^\infty = \langle f, f,... \rangle$ and let $\pi_n, f^\infty = \langle f_1, f_2, ..., f_n, f, f,... \rangle$; f^∞ is called a *stationary policy*. We associate with each $f \in F$ the $S \times 1$ column vector $r(f)$ and the $S \times S$ Markov matrix $Q(f)$ whose sth and (s, s') elements are $i(s, f(s))$ and $q(s' \mid s, f(s))$, respectively. If $\pi = \langle f_i \rangle_{i=1}^{\infty}$, let

$$Q_n(\pi) = Q(f_1)\ Q(f_2) \cdots Q(f_n).$$

Then the vector of total discounted returns from policy π is

$$V_\beta(\pi) = \sum_{n=0}^{\infty} \beta^n Q_n(\pi)\, r(f_{n+1}), \tag{1}$$

where β $(0 \leqslant \beta < 1)$ is the discount factor and $Q_0(\pi) = I$, the $S \times S$ identity matrix.

The vector of total expected returns in periods 1, 2,..., n from policy π is given by

$$V^n(\pi) = \sum_{i=0}^{n-1} Q_i(\pi)\, r(f_{i+1}). \tag{2}$$

Denoting the sth component of a vector u by $[u]_s$, we write $u \geqslant v$ if $[u]_s \geqslant [v]_s$ for all s and $u > v$ if $u \geqslant v$ and $u \neq v$. Also, let $\mathbf{1}$ be the S-vector of 1's.

A policy π^* is called β-optimal $(0 \leqslant \beta < 1)$ or optimal if for all policies π we have

$$V_\beta(\pi^*) \geqslant V_\beta(\pi) \tag{3}$$

or

$$\liminf_{N\to\infty} N^{-1} V^N(\pi^*) \geqslant \limsup_{N\to\infty} N^{-1} V^N(\pi). \tag{4}$$

3. The Set of β-Optimal Policies

Let $A(\beta)$ be the set of β-optimal policies and define the nonempty set (see [1], p. 721) $F(\beta)$ by

$$F(\beta) = \{f \in F : f^\infty \in A(\beta)\}, \qquad 0 \leqslant \beta < 1.$$

Throughout this section we assume that $\beta(0 \leqslant \beta < 1)$ is fixed and that $g \in F(\beta)$.

Lemma 1. *If π is a sequence from $F(\beta)$, then $\pi \in A(\beta)$.*

Proof. We first show by induction that

$$\pi \in A(\beta) \quad \text{if} \quad \pi = \pi_n, g^\infty \quad \text{where} \quad f_i \in F(\beta) \quad \text{for} \quad 1 \leqslant i \leqslant n. \tag{5}$$

For $n = 1$ we have

$$\begin{aligned} V_\beta(\pi) &= r(f_1) + \beta Q(f_1)\, V_\beta(g^\infty) \\ &= r(f_1) + \beta Q(f_1)\, V_\beta(f_1^\infty) \\ &= V_\beta(f_1^\infty) = V_\beta(g^\infty). \end{aligned}$$

Suppose (5) holds for $1 \leqslant i \leqslant n$, then

$$V_\beta(\pi) = \sum_{i=0}^{n-1} \beta^i Q_i(\pi)\, r(f_{i+1}) + \beta^n Q_n(\pi)\,[r(f_{n+1}) + \beta Q(f_{n+1})\, V_\beta(g^\infty)]$$

$$= \sum_{i=0}^{n-1} \beta^i Q_i(\pi)\, r(f_{i+1}) + \beta^n Q_n(\pi)\, V_\beta(g^\infty)$$

$$= V_\beta(g^\infty),$$

the last equality follows from the induction hypothesis.

Now let π be a sequence from $F(\beta)$ and let $\epsilon > 0$ be given. Choose n sufficiently large so that $\beta^n V_\beta(g^\infty) < \epsilon\, \mathbf{1}$. We may assume without loss of generality that $i(s, a) \geqslant 0$. Consequently,

$$| V_\beta(\pi) - V_\beta(g^\infty) | \leqslant | V_\beta(\pi) - V_\beta(\pi_n, g^\infty) | + | V_\beta(\pi_n, g^\infty) - V_\beta(g^\infty) | < \epsilon\, \mathbf{1}$$

by the definition of n and (5). Hence, $\pi \in A(\beta)$ as $\epsilon > 0$ was arbitrary. Q.E.D.

To avoid trivialities, we make the following policy space reduction: if $\sum_{i=1}^{S} [Q_n(\pi)]_{is} = 0$, then we shall assume that $f_n(s) = g(s)$, that is, if the probability under π of entering state s at time n is zero, then the action taken at time n in state s is irrelevant so we may assume that it is $g(s)$.

THEOREM 1. *The set $A(\beta)$ of β-optimal policies is the set of sequences from $F(\beta)$.*

PROOF. Suppose $V_\beta(f, g^\infty) = V_\beta(g^\infty)$, then since

$$V_\beta(f, g^\infty) = r(f) + \beta Q(f)\, V_\beta(g^\infty)$$

we have

$$V_\beta(g^\infty) = [I - \beta Q(f)]^{-1} r(f) = V_\beta(f^\infty),$$

that is, $f \in F(\beta)$.

Take $\pi \in A(\beta)$ and suppose $f_i \in F(\beta)$, $1 \leqslant i \leqslant n$, *and* $f_{n+1} \notin F(\beta)$, then $Q_n(\pi)\, V_\beta(f_{n+1}, g^\infty) < Q_n(\pi)\, V_\beta(g^\infty)$ by the policy space reduction and the above so that

$$V_\beta(\pi) = \sum_{i=0}^{n-1} \beta^i Q_i(\pi)\, r(f_{i+1}) + \beta^n Q_n(\pi)\,[r(f_{n+1}) + \beta Q(f_{n+1})\, V_\beta(f_{n+2}, f_{n+3}, \cdots)]$$

$$\leqslant \sum_{i=0}^{n-1} \beta^i Q_i(\pi)\, r(f_{i+1}) + \beta^n Q_n(\pi)\, V_\beta(f_{n+1}, g^\infty)$$

$$< \sum_{i=0}^{n-1} \beta^i Q_i(\pi)\, r(f_{i+1}) + \beta^n Q_n(\pi)\, V_\beta(g^\infty) = V_\beta(\pi_n, g^\infty).$$

But this contradicts $\pi \in A(\beta)$, so $f_i \in F(\beta)$ for all i.

The converse is provided by Lemma 1. Q.E.D.

Of course, if $F(\beta)$ is a singleton, then g^∞ is the unique β-optimal policy. In particular, there is no β-optimal nonstationary policy.

Since $[V_\beta(f^\infty)]_s = [(I - \beta Q(f))^{-1} r(f)]_s$ for each $f \in F$, we have from Cramer's rule that $[V_\beta(f^\infty)]_s$ is the ratio of two polynomials in β and has at most S roots on $[0, 1)$ (unless $[V_\beta(f^\infty)]_s \equiv 0$). Let $|F|$ be the number of distinct stationary policies. It now follows that $F(\beta)$ is a singleton for all but perhaps $S(2S + 1)(|F| - 1)$ points of $[0, 1)$ if and only if (i) $V_\beta(h^\infty) - V_\beta(f^\infty) \equiv 0$ and $f \neq h$ implies $\{\beta : f^\infty \in F(\beta)\}$ is a finite set. A condition under which (i) holds is (i') $r(f) \neq r(h)$ if $f \neq h$.

The next theorem extends some work of Blackwell [1] and Veinott [3].

THEOREM 2. *There is a finite sequence of pairs* $\langle (a_i, g_i) \rangle_{i=1}^m$ *such that* $0 \equiv a_1 < a_2 < \cdots < a_{m+1} \equiv 1$ *and* g_i *is* β*-optimal for all* $\beta \in [a_i, a_{i+1})$, $i = 1, 2, \ldots, m$.

PROOF. Since $[V_\beta(f^\infty)]_s$ is a rational function, $\{\beta : \beta$ is an isolated root of $[V_\beta(f^\infty) - V_\beta(h^\infty)]_s$, $1 \leqslant s \leqslant S, f, h \in F\}$ is a finite set, enumerated $0 \equiv a_1 < a_2 < \cdots < a_{m+1} \equiv 1$. (We include 0 and 1 whether or not they are roots.) Choose $1 \leqslant i \leqslant m$, $\alpha \in (a_i, a_{i+1})$, and $g_i \in F(\alpha)$, then using Lemma 1, the definition of a_i, a_{i+1}, and the continuity of $[V_\beta(f^\infty)]_s$ for $0 \leqslant \beta < 1$ $(1 \leqslant s \leqslant S, f \in F)$, it follows that g_i^∞ is β-optimal for all $\beta \in [a_i, a_{i+1})$. Q.E.D.

4. THE SET OF POLICIES WITH MAXIMAL RETURN PER UNIT TIME

Let $\bar{A}$ be the set of optimal policies (see (4)), and define the set $\bar{F}$ by

$$\bar{F} = \{f \in F : f^\infty \in \bar{A}\}.$$

By Brown's Theorem 4.2 [2], $\max_\pi V^n(\pi) - V^n(f^\infty)$ is bounded uniformly in n for some $f \in F$, say g, so it follows that $\bar{F}$ is nonempty. Of course, $g^\infty \in \bar{A}$.

THEOREM 3. *If* $\pi \in \bar{A}$, *then*

$$Q_n(\pi)\, x(g) = x(g), \qquad \text{all} \quad n, \tag{6}$$

where

$$x(g) = \lim_{N \to \infty} N^{-1} V^N(g^\infty).$$

PROOF. Since $\max_\tau V^n(\tau) - V^n(g^\infty)$ is bounded uniformly in n, we have for each fixed n

$$x(g) \geqslant Q_n(\pi)\, x(g) \tag{7}$$

since

$$\begin{aligned} x(g) &= \limsup_{N\to\infty} N^{-1} \max_{\tau} V^N(\tau) \\ &\geqslant \limsup_{N\to\infty} N^{-1} V^N(\pi_n, g^\infty) \\ &= \lim_{N\to\infty} N^{-1} Q_n(\pi)\, V^N(g^\infty) \\ &= Q_n(\pi)\, x(g). \end{aligned}$$

Since $\pi \in \bar{A}$, we have from (7) that for each fixed n

$$\begin{aligned} x(g) &= \limsup_{N\to\infty} N^{-1} V^N(\pi) = Q_n(\pi) \limsup_{N\to\infty} N^{-1} V^N(f_{n+1}, f_{n+2}, \ldots) \\ &\leqslant Q_n(\pi)\, x(g) \leqslant x(g). \end{aligned}$$

Q.E.D.

COROLLARY 1. *If* $\pi = \langle f_1, f_2, \ldots, f_n, g^\infty \rangle$ *and* $f_i \in \bar{F}$ *for* $1 \leqslant i \leqslant n$, *then* $\pi \in \bar{A}$.

PROOF. Applying (6) to $f_1, f_2, \ldots, f_n$ in reverse order we have

$$\begin{aligned} \liminf_{N\to\infty} N^{-1} V^N(\pi) &= Q_n(\pi) \lim_{N\to\infty} N^{-1} V^N(g^\infty) \\ &= Q_n(\pi)\, x(g) = Q_{n-1}(\pi) Q(f_n)\, x(g) \\ &= Q_{n-1}(\pi)\, x(g) = \cdots = x(g). \end{aligned}$$

Q.E.D.

Unlike Theorem 1, however, it is not true that all sequences from $\bar{F}$ are in $\bar{A}$. Nor is it true that each policy in $\bar{A}$ is a sequence from $\bar{F}$. This is shown in examples 1 and 2, respectively.

EXAMPLE 1. *A sequence from* $\bar{F}$ *which is not in* $\bar{A}$. Let $S = 3$, $A_i = \{1, 2\}$, $i = 1, 2, 3$,

$$\begin{aligned} q(1 \mid 1, 1) = q(1 \mid 2, 1) = q(2 \mid 3, 1) = q(3 \mid 3, 2) = q(3 \mid 2, 2) \\ = q(2 \mid 1, 2) = 1, \end{aligned}$$

and

$$i(s, a) = \begin{cases} 1, & \text{if} \quad a = s = 1 \quad \text{or if} \quad a = 2 \quad \text{and} \quad s = 3 \\ 0, & \text{otherwise.} \end{cases}$$

If $f = 1$ and $g = 21$, then $x(f) = x(g) = 1$; yet,

$$\lim_{N\to\infty} N^{-1} V^N(\pi) = 01,$$

where $\pi = \langle f, g, f, g, \cdots \rangle$.

EXAMPLE 2. *A policy in $\bar{A}$ which is not a sequence from $\bar{F}$.* Let $S = 1$, $A_1 = \{1, 2\}$, $i(1, 1) = 1$, $i(1, 2) = 2$, $f = 1$ and $g = 2$. Then

$$x(f) = 1 < 2 = x(g),$$

but

$$\lim_{N\to\infty} N^{-1}V^N(\pi) = 2,$$

where $\pi = \langle f_i\rangle_{i=1}^{\infty}$ is defined by

$$f_i = \begin{cases} f, & \text{if} \quad i = 2^n \quad \text{for} \quad n = 1, 2,..., \\ g, & \text{otherwise.} \end{cases}$$

In fact, $f_i = f \notin \bar{F}$ for infinitely many i.

REFERENCES

1. D. BLACKWELL. Discrete dynamic programming. *Ann. Math. Stat.*, **33** (1962), 719–726.
2. B. W. BROWN. On the iterative method of dynamic programming on a finite space discrete time Markov process. *Ann. Math. Stat.* **36** (1965), 1279–1285.
3. A. F. VEINOTT. JR. On finding optimal policies in discrete dynamic programming with no discounting. *Ann. Math. Stat.* **37** (1966), 1284–1294.